VUB SCIENCES & BIOENGINEERING SCIENCES

The Research Group
Artificial Intelligence Lab

has the honor to invite you to the public defence of the PhD thesis of

Willem Röpke

to obtain the degree of Doctor of Sciences

Title of the PhD thesis:

Thinking in Trade-Offs: Training Agents to Balance Conflicting Objectives Under Uncertainty

Supervisor:
Prof. dr. Ann Nowé (VUB)
Co-supervisor:
Prof. dr. Roxana Rădulescu (VUB, Universiteit Utrecht, NL)
Dr. Diederik M. Roijers (VUB, gemeente Amsterdam, NL)

The defence will take place on
Wednesday, February 18, 2026 at 4 p.m.

VUB Etterbeek campus, Pleinlaan 2, Elsene,
In auditorium I.0.03

Members of the jury

Prof. dr. Beat Signer (VUB, chair)
Prof. dr. Bart Bogaerts (VUB)
Prof. dr. Marie-Anne Guerry (VUB)
Prof. dr. Guillermo Alberto Pérez (UAntwerpen)
Prof. dr. Giorgia Ramponi (University of Zürich, CH)

Curriculum vitae

Willem Röpke obtained an FWO fellowship and joined the AI Lab at Vrije Universiteit Brussel in 2021 to pursue his PhD on multi-objective reinforcement learning. He has published in leading journals and conferences, and has made contributions to open-source projects. He was a visiting researcher at the University of Oxford and the University of Galway, served as a teaching assistant in machine learning, and has organised workshops including MODeM 2024 and EWRL 2023.

Abstract of the PhD research

Every day, people face decisions that involve multiple, often conflicting objectives: balancing cost against quality, safety against speed, or personal benefit against social responsibility. These trade-offs rarely admit a single “right”' answer, and they become even more challenging when other decision-makers are involved. Reinforcement learning offers a powerful framework for constructing artificial agents that act autonomously in complex, uncertain environments, learning through trial and error how to make effective decisions. Yet, most existing approaches focus on a single objective, assuming away the trade-offs that are intrinsic to real decision-making. This thesis addresses that gap directly. Its central theme is how to design agents that can *think in trade-offs*, that is, agents that can reason about multiple objectives and act optimally under uncertainty.

The contributions unfold in three parts. First, we bridge single- and multi-objective reinforcement learning by showing how decomposition techniques allow well-established single-objective methods to be extended to learning the Pareto front, a classical solution set that captures efficient trade-offs for certain decision-makers. Building on this foundation, we then introduce and analyse alternative solution concepts that more directly reflect decision-makers' preferences, developing rigorous theoretical guarantees and demonstrating their practical relevance. Finally, we turn to multi-agent systems and establish a novel reduction from multi-objective to single-objective games, which not only provides new theoretical insights but also enables the transfer of powerful algorithms across domains.

By establishing strong connections between multi-objective and single-objective paradigms, the thesis lays the groundwork for future progress to be accelerated, enabling advances in one field to be immediately translated into the other. These advances bring us closer to systems that can adapt their behaviour to different stakeholders, balance conflicting objectives transparently, and operate responsibly in safety-critical real-world environments.